PB84-175280

DISFIT
A Program for Fitting Distributions in Data

(U.S.) Health Effects Research Lab.
Research Triangle Park, NC

1981

U.S. Department of Commerce
National Technical Information Service
NTIS

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before comple

1. REPORT NO. EPA-600/J-81-688	2.
3 PB84-175280	
4. TITLE AND SUBTITLE DISFIT: A Program for Fitting Distributions in Data (Journal Version)	5. REPORT DATE 1981
	6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) Victor Hasselblad and Andrew Stead	8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS Health Effects Research Laboratory US Environmental Protection Agency Research Triangle Park, NC 27711	10. PROGRAM ELEMENT NO.
	11. CONTRACT/GRANT NO.
12. SPONSORING AGENCY NAME AND ADDRESS Office of Research & Development Health Effects Research Laboratory US Environmental Protection Agency Research Triangle Park, NC 27711	13. TYPE OF REPORT AND PERIOD COVERED
	14. SPONSORING AGENCY CODE EPA-600/11

15. SUPPLEMENTARY NOTES

REF: The American Statistician, 35(4):274-279, November 1981

16. ABSTRACT

Although distribution fitting methods abound in the statistical literature, very few of these methods are found in the major statistical packages. In particular, SPSS (1975), BMD-P (1981) and SAS (1979) only give some overall tests for normality. There a few specialized distribution fitting packages. Bouver and Bargmann (1975) wrote a system to fit the Pearson system of curves by the method of moments. The system of Alldredge and Bolstad (1977) fits the normal, lognormal, exponential, gamma, and uniform distributions using maximum likelihood for some distributions and moment estimators for others. Holland and Fitz-Simons (1982) fitted the normal three parameter lognormal, three parameter Weibull, three parameter gamma, Johnson S_B, and four parameter beta. Their system obtains maximum likelihood estimates and uses the Nelder-Mead Simplex iterative search method when closed solutions are not available. Morgan, Nelson, and Caporal (1979) have written a system that includes the normal, lognormal, Weibull and binomial. This system has the advantage that linear models can be fitted to the shape parameter, using maximum likelihood methods. This feature limits the kinds of distributions which can be fitted, however. The distributions we have chosen for DISFIT include some of the classic ones as well as those pertinent ot the environmental field. The analysis of both environmental monitoring and environmental health data require distribution fitting programs.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group

18. DISTRIBUTION STATEMENT	19. SECURITY CLASS (This Report)	21. NO. OF PAGES 7
	20. SECURITY CLASS (This page)	22. PRICE

EPA Form 2220-1 (Rev. 4-77) PREVIOUS EDITION IS OBSOLETE

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

EPA-600/J-81-688
JOURNAL ARTICLE

DISFIT: A PROGRAM FOR FITTING DISTRIBUTIONS TO DATA

Victor Hasselblad and Andrew G. Stead, Environmental Protection Agency

1. INTRODUCTION

Although distribution fitting methods abound in the statistical literature, very few of these methods are found in the major statistical packages. In particular, SPSS (1975), BMD-P (1981) and SAS (1979) only give some overall tests for normality. There are a few specialized distribution fitting packages. Bouver and Bargmann (1975) wrote a system to fit the Pearson system of curves by the method of moments. The system of Alldredge and Bolstad (1977) fits the normal, lognormal, exponential, gamma, and uniform distributions using maximum likelihood for some distributions, and moment estimators for others. Holland and Fitz-Simons (1982) fitted the normal, three parameter lognormal, three parameter Weibull, three parameter gamma, Johnson S_B, and four parameter beta. Their system obtains maximum likelihood estimates and uses the Nelder-Mead Simplex iterative search method when closed solutions are not available. Morgan, Nelson, and Caporal (1979) have written a system that includes the normal, lognormal, Weibull and binomial. This system has the advantage that linear models can be fitted to the shape parameter, using maximum likelihood methods. This feature limits the kinds of distributions which can be fitted, however.

The distributions we have chosen for DISFIT include some of the classic ones as well as those pertinent to the environmental field. The analysis of both environmental monitoring and environmental health data require distribution fitting programs. For example, ambient air standards are often based on the distribution of the "second maximum". Health surveys often yield counts, and the distributional form of these counts may be important for hypothesis testing.

The DISFIT system is written and programmed in three separate parts. The first part fits discrete distributions, and is available as a technical report (Hasselblad, Stead and Anderson, 1981). The second part fits continuous distributions. This program is available, and the report describing the program should be available by the end of 1982. The third part will fit grouped continuous distributions.

Much of the information on these distributions can be found in the three volume work by Johnson and Kotz (1969, 1970a, 1970b). This classic set is so comprehensive, that only a small portion of its information is contained in our reports. Each section of our report discusses a particular distribution, and includes a definition of the distribution, equations for estimating parameters, and actual examples. Parameter estimates and their variances are given, but tests of hypotheses are not provided. Tests often can be made from the likelihood functions, that are provided.

The routines are written so that each parameter may be specified or estimated. In some cases, specifying a parameter value may give rise to another distribution. For example, the Weibull distribution with C set to 1 becomes the exponential.

2. METHODS OF ESTIMATION

Estimates in the DISFIT system are calculated by maximum likelihood. There are, of course, several reasons for using maximum likelihood estimates. The estimates have many optimal properties including asymptotic consistency and efficiency (Kendall and Stuart, 1961). The comparison of the likelihood function for different distributions can give some measure of goodness-of-fit. In some specific cases, it is possible to test the fit of one distribution versus another using likelihood ratio tests.

Maximum likelihood estimates also have certain problems. Some of these are caused by small sample sizes, although distribution fitting usually requires moderate to large sample sizes. The likelihood function for mixtures of continuous distributions can have singularities corresponding to each data point, as shown by Day (1969). Also, the likelihood function may have more than one local maximum. We can only hope that the initial estimates lead to meaningful solutions in each case.

Distributions such as the gamma or beta, or functions of them such as the negative binomial, require iterative methods to obtain maximum likelihood estimates. For these cases, the modified Gauss-Newton method is easy to compute, and has the advantage of being a second order method. A description of the method was given by Hartley (1961). In other cases we have used the standard Newton-Raphson method.

Maximum likelihood estimation is especially useful for problems of "missing information". Missing information can take the form of censored data, truncated distributions, grouping, or mixing of distributions. All of these problems are handled by using the "missing information principle" (MIP) of Orchard and Woodbury (1972). This method will converge geometrically under a wide variety of conditions, as shown by Dempster, Laird, and Rubin (1977).

3. EXAMPLES OF DISCRETE DISTRIBUTIONS

The discrete portion of the DISFIT system contains routines to fit the binomial, truncated binomial, mixture of two binomials, beta binomial, Poisson, truncated Poisson, mixture of two Poissons, negative binomial, truncated negative binomial, and logarithmic distributions. Output includes estimates of the distributional parameters (if not specified), asymptotic variances (or covariances), and log-likelihood function, as well as a chi-square goodness-of-fit test.

3.1 Example of a binomial distribution

Fisher (1958) gives counts from a dice experiment of Weldon. The counts represent the number of 5's and 6's occurring in the toss of 12 dice. The distribution of this number should be given by a binomial distribution where $p = 1/3$. The estimate of p from the observed data is .3377, as shown in Figure 3.1. A likelihood ratio test of $p = 1/3$ yields an asymptotic chi-square of $-2(-50255.16 + 50241.65) = 27.02$ ($p<.0001$), indicating that $p = .3377$ provides a significantly better fit.

The American Statistician, 35(4):274-279, Nov. 1981
Reprinted from the 1982 Statistical Computing Section Proceedings of the American Statistical Association

```
WELDON'S DICE DATA

FIT TO A BINOMIAL DISTRIBUTION

P(X,N) = N!P**X(1-P)**(N-X)/(X!(N-X)!)

TOTAL SAMPLE SIZE = 26306, N =  12

ESTIMATED VALUE OF P = .3377
VARIANCE OF P =    .70851-006

HOMOGENEITY TEST = 26445.79, P = .2691

   CHI-SQUARE GOODNESS OF FIT TEST

VALUE  OBSERVED  EXPECTED  CHI-SQUARE

  0       185      187.38        .03
  1      1149     1146.51        .01
  2      3265     3215.24        .77
  3      5475     5464.70        .02
  4      6114     6269.35       3.85
  5      5194     5114.65       1.23
  6      3067     3042.54        .20
  7      1331     1329.73        .00
  8       403      423.76       1.02
  9       105       96.03        .84
 10        14       14.69        .22
 11         4        1.36         --
 12         0         .06         --

CHI-SQUARE =    8.18, D.F. =  9, P = .5161

LOG-LIKELIHOOD  = -50241.6543
```

Figure 3.1

3.2 Example of betabinomial distribution

Haseman and Soares (1976) fitted a beta-binomial distribution to investigate fetal mortality in pregnant mice. The N of the distribution is the number of fetuses per litter; p is the proportion of dead fetuses/litter. Some 524 mice were tested, where N ranged from 1 to 20. Figure 3.2 gives the fit for an N of 13, which is adequate as measured by the chi-square test.

```
HASEMAN AND SOARES DATA FOR N = 13

FIT TO A BETA-BINOMIAL DISTRIBUTION

P(X) = N!(A+X-1)!(B+N-X-1)!(A-1)!(B-1)!/
       (X!(N-X)!(A+B+N-1)!(A+B-1)!

TOTAL SAMPLE SIZE =    92, N =  13

ESTIMATED VALUE OF ALPHA =      1.03823
ESTIMATED VALUE OF BETA  =     10.53389

ASYMPTOTIC COVARIANCE MATRIX OF ESTIMATES
            ALPHA          BETA
ALPHA    .10745+000    .10081+001
BETA     .10081+001    .11536+002

   CHI-SQUARE GOODNESS OF FIT TEST

VALUE  OBSERVED  EXPECTED  CHI-SQUARE

  0        39       39.97        .02
  1        27       23.94        .39
  2        12       13.60        .19
  3         6        7.38        .26
  4         5        3.81        .12
  5         2        1.87         --
  6         0         .86         --
  7         0         .36         --
  8         1         .14         --
 9+         0         .05         --

CHI-SQUARE =     .98, D.F. =  2, P = .6140

LOG-LIKELIHOOD =   -137.1281
```

Figure 3.2

3.3 Example of a Poisson distribution

Mutagenicity testing using the Ames test (1975) has recently generated a great deal of interest. The test consists of counts of revertants per plate for each dose of a given compounds. The following example (Figure 3.3) is the result of 196 plates at a zero dose (controls) for the Salmonella strain TA1537 without metabolic activation. The plates were all made on the same day. After a 48-hour incubation period revertants per plate were counted manually. The fit is adequate as measured by both the variance test and the standard chi-square goodness-of-fit test.

```
REVERTANT COUNTS FOR STRAIN TA1537 - MANUAL

FIT TO A POISSON DISTRIBUTION

P(X) = EXP(-LAMBDA)LAMBDA**X/X!

TOTAL SAMPLE SIZE =   196

ESTIMATED VALUE OF LAMBDA =  10.9082
VARIANCE OF LAMBDA =    .55654-001

VARIANCE TEST =  205.20, P = .2941

   CHI-SQUARE GOODNESS OF FIT TEST

VALUE  OBSERVED  EXPECTED  CHI-SQUARE

  0         0        .00         --
  1         0        .04         --
  2         0        .21         --
  3         1        .78         --
  4         3       2.12         --
  5         4       4.62        .01
  6         6       8.40        .68
  7        18      13.08       1.85
  8        20      17.84        .26
  9        23      21.62        .09
 10        19      23.59        .89
 11        18      23.39       1.24
 12        22      21.26        .03
 13        14      17.84        .83
 14        19      13.90       1.87
 15        11      10.11        .03
 16         9       6.89        .65
 17         4       4.42        .04
 18         1       2.68        .14
 19         3       1.54         --
 20         0        .84         --
 21         0        .44         --
 22         1        .22         --
 23+        0        .18         --

CHI-SQUARE =    8.64, D.F. = 12, P = .7330

LOG-LIKELIHOOD  =   -515.1158
```

Figure 3.3

3.4 Example of a mixture of two Poisson distributions

A mixture of two Poisson distributions can be thought of as a Poisson distribution where the parameter λ follows a binomial distribution. Equations for obtaining maximum likelihood estimates were given by Hasselblad (1969). An example of a possible mixture of two Poisson distributions was found by pooling control Ames test data, strain TA1537, from several experiments. Such a distribution could arise if a change occurred in either the bacterial strain or the laboratory technique during the coarse of the experiments. The pooled data are shown in Figure 3.4. The log-likelihood for a single Poisson distribution is -724.01 and for a mixture of two Poissons is -674.88. The likelihood ratio test for the adequacy of a single Poisson distribution gives a chi-square of 98.26 for 2 degrees of freedom ($p<.001$), indicating the

mixture of 2 Poissons provides a significantly better fit.

```
COUNTS OF REVERTANTS: TA1537-

FIT TO A MIXTURE OF TWO POISSON DISTRIBUTIONS

P(X) = ALPHA*EXP(-LAMBDA1)X**LAMBDA1/X! +
       (1-ALPHA)EXP(-LAMBDA2)X**LAMBDA2)/X!

TOTAL SAMPLE SIZE =   238

ESTIMATED VALUE OF LAMBDA1 =     6.49579
ESTIMATED VALUE OF LAMBDA2 =    13.28531
ESTIMATED VALUE OF ALPHA =        .68377

ASYMPTOTIC COVARIANCE MATRIX OF ESTIMATES
          LAMBDA1        LAMBDA2        ALPHA
LAMBDA1  .13070+000     .18212+000     .17625-001
LAMBDA2  .18212+000     .64116+000     .39984-001
ALPHA    .17625-001     .39984-001     .45459-002

   CHI-SQUARE GOODNESS OF FIT TEST

VALUE  OBSERVED  EXPECTED  CHI-SQUARE

  0        0        .25        --
  1        1       1.60        --
  2        6       5.19        .00
  3       13      11.27        .26
  4       21      18.39        .37
  5       24      24.12        .00
  6       22      26.61        .80
  7       23      25.64        .27
  8       22      22.40        .01
  9       20      18.49        .12
 10       20      15.09       1.60
 11        6      12.64       3.49
 12       17      10.97       3.32
 13       11       9.70        .18
 14       11       8.50        .74
 15        4       7.22       1.44
 16        5       5.87        .13
 17        3       4.54        .52
 18        4       3.34        .13
 19        0       2.33        .22
 20        0       1.54        --
 21        1        .98        --
 22        0        .59        --
 23        2        .34        --
 24        0        .19        --
 25        0        .10        --
 26        0        .05        --
 27        1        .03        --
 28        1        .01        --
29+        0        .01        --

CHI-SQUARE =   13.60, D.F. = 14, P = .4802

LOG-LIKELIHOOD =   -674.8829
```

Figure 3.4

4. EXAMPLES OF CONTINUOUS DISTRIBUTIONS

The continuous portion of the DISFIT System contains routines to fit the normal, truncated normal, mixture of two normals, lognormal, truncated lognormal, mixture of two lognormals, gamma, Weibull, logistic, exponential, mixture of two exponentials, and beta distributions. The observations are read in as the actual values, rather than as counts of observations from intervals. If the data are in the form of counts, then a grouped data program should be used. Censored distributions are a special case of grouped distributions, and will also be handled by the grouped data portion of the DISFIT system. In order to compute goodness-of-fit statistics, interval boundaries are read in by the continuous portion of the DISFIT system. The output shows counts in intervals, but the actual values are used in the computations.

4.1 Example of a normal distribution

The normal distribution has received more attention than any other distribution. Although appropriately transformed data and residuals are often fitted well by a normal distribution, we found few examples of raw data so distributed. Ali and Giaccotto (1982) give monthly log-price changes for several stocks over the period January 1, 1966 to December 31, 1976. In a sense the data have already been transformed, but the log-price change is a meaningful measurement. AT&T stock provides a good example of normally distributed log-price change data. The fit is shown in figure 4.1.

```
MONTHLY LOG-PRICE CHANGES FROM 1/1/66 TO 12/31/76 FOR AT&T

FIT TO A NORMAL DISTRIBUTION

F(X) = EXP((X - MU)**2/2*SIGMA**2)/SIGMA*SQRT(2*PI)

ESTIMATED VALUE OF MU =              .004852
ESTIMATED VALUE OF SIGMA =           .044605
UNBIASED ESTIMATE OF SIGMA**2 =      .002005

ESTIMATED VARIANCE OF MU =        .15073-004
ESTIMATED VARIANCE OF SIGMA =     .75150-005

FISHER'S G STATISTICS

            VALUE       VARIANCE       P-VALUE

G1 =       .03876        .044458       .854141
G2 =      -.06482        .175283       .876969

LOG-LIKELIHOOD =        223.2070

     ESTIMATED PERCENTILES FROM THE DISTRIBUTION

PERCENTILE  VALUE   PERCENTILE  VALUE   PERCENTILE  VALUE

    .1    -.1330      10.0    -.0523      95.0      .0782
    .5    -.1100      25.0    -.0252      97.5      .0923
   1.0    -.0989      50.0     .0049      99.0      .1086
   2.5    -.0826      75.0     .0349      99.5      .1197
   5.0    -.0685      90.0     .0620      99.9      .1427

          CHI-SQUARE GOODNESS OF FIT TEST

      INTERVAL              OBSERVED  EXPECTED  CHI-SQUARE

          TO     -.1000        2        1.24        --
-.1000 TO        -.0750        3        3.61        --
-.0750 TO        -.0500        8        9.60        .14
-.0500 TO        -.0250       19       18.78        .00
-.0250 TO         .0000       23       27.06        .61
 .0000 TO         .0250       35       28.72       3.00
 .0250 TO         .0500       18       22.44        .88
 .0500 TO         .0750       12       16.91        .06
 .0750 TO         .1000        7        5.47        .24
 .1000 TO                      2        2.17        --

CHI-SQUARE =    4.94, D.F. = 4, P = .2934
```

Figure 4.1

4.2 Example of a truncated normal distribution

Kenyon, Scheffer and Chapman (1950) published a fascinating survey of the Alaskan fur-seal herd. One of their many tables contains lengths of 3-year-old male seals recovered on St. Paul Island (in the Pribilof Islands). The seals included in this analysis were those with tags lost. The sealers are under instructions to kill only 41 to 45-inch males. There are very few 3-year old males over 45 inches long, but many are under 41 inches. Since 40.5 inches was the intended truncation point, some 18 40-inch and one 39-inch seals were excluded from the fit. The truncation over 45 inches was ignored in this example. The fit is shown in figure 4.2. Although the fit is not perfect, the data represent the only real example of a truncated normal distribution we found in the literature.

LENGTHS OF SEALS RECOVERED ON ST. PAUL ISLAND IN 1950

FIT TO A LEFT TRUNCATED NORMAL DISTRIBUTION

TRUNCATION POINT = 40.50000

ESTIMATED VALUE OF MU = 41.880774
ESTIMATED VALUE OF SIGMA = 1.482148

ASYMPTOTIC COVARIANCE MATRIX OF PARAMETER ESTIMATES

	MU	SIGMA
MU	.82555-002	.20008-002
SIGMA	.20008-002	.34714-002

LOG-LIKELIHOOD = -580.9383

ESTIMATED PERCENTILES FROM THE DISTRIBUTION

PERCENTILE	VALUE	PERCENTILE	VALUE	PERCENTILE	VALUE
.1	40.5047	10.0	40.9190	95.0	44.4550
.5	40.5235	25.0	41.4351	97.5	44.9064
1.0	40.5466	50.0	42.2100	99.0	45.4350
2.5	40.6141	75.0	43.0964	99.5	45.7966
5.0	40.7211	90.0	43.9394	99.9	46.5454

CHI-SQUARE GOODNESS OF FIT TEST

INTERVAL		OBSERVED	EXPECTED	CHI-SQUARE
TO	41.5000	109	110.31	.02
41.5000 TO	42.5000	144	130.35	1.43
42.5000 TO	43.5000	81	99.37	3.40
43.5000 TO	44.5000	55	48.86	.77
44.5000 TO	45.5000	17	15.49	.00
45.5000 TO		2	3.62	--

CHI-SQUARE = 5.61, D.F. = 2, P = .0604

Figure 4.2

4.3 Example of a lognormal distribution

Tepper and Levin (1975) reported on a study of blood lead levels from housewives in seven U.S. metropolitan areas. Although there was significant variation from area to area, the largest portion of variation was between individuals. A lognormal curve was fitted to the entire data set consisting of 2102 blood lead values. All values 10 or below were pooled in the goodness-of-fit, since the measurement error is very large in this range. The results are shown in figure 4.3.

BLOOD LEAD VALUES FROM THE SEVEN CITY STUDY

FIT TO A LOGNORMAL DISTRIBUTION

F(X) = EXP(-(LOG(X/MU))**2/2*SIGMA**2)/(X*SIGMA*SQRT(2PI))

ESTIMATED VALUE OF MU = 2.788927
ESTIMATED VALUE OF SIGMA = .297505

ESTIMATED VARIANCE OF MU = .42107-004
ESTIMATED VARIANCE OF SIGMA = .21051-004

GEOMETRIC MEAN = 16.26356, GEOM. STD. DEV. = 1.34649

LOG-LIKELIHOOD = -6296.6251

ESTIMATED PERCENTILES FROM THE DISTRIBUTION

PERCENTILE	VALUE	PERCENTILE	VALUE	PERCENTILE	VALUE
.1	6.4855	10.0	11.1079	95.0	26.5301
.5	7.5530	25.0	13.3066	97.5	29.1375
1.0	8.1403	50.0	16.2636	99.0	32.4930
2.5	9.0778	75.0	19.8775	99.5	34.9965
5.0	9.9700	90.0	23.8121	99.9	40.7840

CHI-SQUARE GOODNESS OF FIT TEST

INTERVAL		OBSERVED	EXPECTED	CHI-SQUARE
TO	10.5000	139	148.57	.62
10.5000 TO	13.5000	399	409.84	.29
13.5000 TO	16.5000	550	533.25	1.34
16.5000 TO	19.5000	445	440.86	.04
19.5000 TO	22.5000	265	280.18	.82
22.5000 TO	25.5000	160	152.04	.42
25.5000 TO	28.5000	78	74.63	.13
28.5000 TO	31.5000	30	34.75	.65
31.5000 TO	34.5000	17	15.55	.13
34.5000 TO	37.5000	3	6.83	2.14
37.5000 TO	40.5000	4	2.96	.11
40.5000 TO		2	2.27	--

CHI-SQUARE = 6.69, D.F. = 8, P = .5703

Figure 4.3

4.4 Example of a gamma distribution

Both the two and three parameter gamma distribution can easily be fitted using the modified Gauss-Newton method. Birnbaum and Saunders (1958) published life-length data which can adequately be described by a two-parameter gamma distribution. The data are lifetimes of 6061-T6 aluminum sheeting under periodic loading maximum stress of 21,000 psi, at 18 cycles per second. The data are shown in figure 4.4.

LIFETIMES OF ALUMINUM SHEETING (BIRNBAUM AND SAUNDERS)

FIT TO A 2 PARAMETER GAMMA DISTRIBUTION

F(X) = X**(ALPHA-1)EXP(-X/BETA)/(BETA**ALPHA)(GAMMA(ALPHA))

ESTIMATED VALUE OF ALPHA = 11.855841
ESTIMATED VALUE OF BETA = 118.162856

ASYMPTOTIC COVARIANCE MATRIX OF PARAMETER ESTIMATES

	ALPHA	BETA
ALPHA	.20869+001	-.22192+002
BETA	-.22192+002	.24872+003

LOG-LIKELIHOOD = -747.2706

ESTIMATED PERCENTILES FROM THE DISTRIBUTION

PERCENTILE	VALUE	PERCENTILE	VALUE	PERCENTILE	VALUE
.1	468.1419	10.0	911.3016	95.0	2129.8659
.5	573.3860	25.0	1109.4085	97.5	2303.9154
1.0	630.1132	50.0	1361.7733	99.0	2517.9102
2.5	720.5021	75.0	1649.3461	99.5	2670.9633
5.0	805.2425	90.0	1940.3460	99.9	3005.8569

CHI-SQUARE GOODNESS OF FIT TEST

INTERVAL		OBSERVED	EXPECTED	CHI-SQUARE
TO	800.0000	6	4.86	--
800.0000 TO	1000.0000	9	11.18	.07
1000.0000 TO	1200.0000	15	17.97	.49
1200.0000 TO	1400.0000	20	20.31	.00
1400.0000 TO	1600.0000	21	17.78	.59
1600.0000 TO	1800.0000	14	12.60	.11
1800.0000 TO	2000.0000	10	7.92	.55
2000.0000 TO	2200.0000	3	4.34	.58
2200.0000 TO	2400.0000	2	2.16	--
2400.0000 TO		1	1.69	--

CHI-SQUARE = 2.39, D.F. = 4, P = .6649

Figure 4.4

4.5 Example of a beta distribution

The beta distribution usually arises from the distribution of certain test statistics. For example, an F distribution can be transformed into a beta distribution. Certain measurements, however, are bounded both below and above, and therefore have the potential to be fitted by a beta distribution. Two of these are percent of cloud cover and percent relative humidity. The following relative humidity measurements were taken from the Raleigh-Durham, N.C. airport at 10:00 AM for the calendar year 1979. 54 days with measurable precipitation were deleted from the data set. The remaining values were fitted to a beta distribution as shown in figure 4.5. The fit is not especially good, but it is an interesting example.

PERCENT RELATIVE HUMIDITY DATA AT RDU DURING 1979 AT 10:00

FIT TO A BETA DISTRIBUTION

F(X) = (X-A)**(P-1)*(B-X)**(Q-1)/(B-A)**(P+Q-1)*BETA(P,Q)
WHERE A = .00000, B = 100.00000

ESTIMATED VALUE OF P = 6.187569
ESTIMATED VALUE OF Q = 3.033257

ASYMPTOTIC COVARIANCE MATRIX OF PARAMETER ESTIMATES

	P	Q
P	.25456+000	.86638-001
Q	.86638-001	.39261-001

LOG-LIKELIHOOD = -1267.3766

ESTIMATED PERCENTILES FROM THE DISTRIBUTION

PERCENTILE	VALUE	PERCENTILE	VALUE	PERCENTILE	VALUE
.1	20.0387	10.0	46.9132	95.0	88.9754
.5	26.5984	25.0	57.2890	97.5	91.5255
1.0	30.1455	50.0	68.3865	99.0	93.9335
2.5	35.7275	75.0	78.2113	99.5	95.2566
5.0	40.8172	90.0	85.4583	99.9	97.2829

CHI-SQUARE GOODNESS OF FIT TEST

INTERVAL		OBSERVED	EXPECTED	CHI-SQUARE
TO	29.5000	2	2.76	--
29.5000 TO	34.5000	1	3.69	1.85
34.5000 TO	39.5000	5	6.69	.42
39.5000 TO	44.5000	10	10.87	.07
44.5000 TO	49.5000	18	16.19	.20
49.5000 TO	54.5000	27	22.32	.98
54.5000 TO	59.5000	29	28.69	.00
59.5000 TO	64.5000	44	34.48	2.63
64.5000 TO	69.5000	42	38.66	.29
69.5000 TO	74.5000	44	40.18	.36
74.5000 TO	79.5000	29	33.13	2.19
79.5000 TO	84.5000	23	32.09	2.58
84.5000 TO	89.5000	16	22.54	1.90
89.5000 TO	94.5000	17	11.34	3.88
94.5000 TO		4	2.36	--

CHI-SQUARE = 17.35, D.F. = 10, P = .0669

Figure 4.5

5. EXAMPLES OF GROUPED CONTINUOUS DISTRIBUTIONS

The grouped continuous portion of the DISFIT system has not been written. Two routines, the grouped (log)normal and the grouped mixture of two lognormal distributions, were written for the analysis of blood lead values as described by Hasselblad, Stead, and Galke (1980). Two examples from this data set are given below. The estimates were obtained using the EM algorithm.

5.1 Example of a grouped lognormal distribution

Billick, Curran, and Shier (1979) described a data set of blood lead values measured in children in New York City as part of a large-scale screening program. The raw values were not recorded; instead the data were grouped into intervals: 1-14, 15-24, 25-34, and so on. Figure 5.1 gives the counts of black children aged 13 to 48 months for the year 1974. Note that the lognormal distribution does a good job of describing these 2236 blood lead values.

5.2 Example of a grouped mixture of two lognormal distributions

Another subset of the data described in example 5.1 provides an example of a mixture of lognormals. While the 1974 blood lead data fit well to a single lognormal, similar data from 1972 did not (χ^2_6 = 80.04, p < .001). From figure 5.2 it can be seen, however, that the

BILLICK BLOOD LEAD DATA - BLACKS, AGE 13 - 48 MONTHS, 1974

FIT TO A GROUPED LOGNORMAL DISTRIBUTION

ESTIMATED VALUE OF MU = 3.215000
ESTIMATED VALUE OF SIGMA = .392337

GEOMETRIC MEAN = 24.903281, GEOM. STD. DEV. = 1.48044

LOG OF LIKELIHOOD = -3260.6244

ASYMPTOTIC COVARIANCE MATRIX OF PARAMETER ESTIMATES

	MU	SIGMA
MU	.76358-004	-.47038-005
SIGMA	-.47038-005	.43321-004

EFFICIENCY OF MEAN COMPARED TO UNGROUPED SAMPLE =90.16%

ESTIMATED PERCENTILES FROM THE DISTRIBUTION

PERCENTILE	VALUE	PERCENTILE	VALUE	PERCENTILE	VALUE
.1	7.4089	30.0	20.2755	80.0	34.6517
1.0	9.9984	40.0	22.5501	90.0	41.1809
5.0	13.0607	50.0	24.9033	95.0	47.4840
10.0	15.0597	60.0	27.5021	99.0	62.0273
20.0	17.8973	70.0	30.5873	99.9	83.7067

CHI-SQUARE GOODNESS OF FIT TEST

INTERVAL		OBSERVED	EXPECTED	CHI-SQUARE
TO	14.5000	186	187.87	.02
14.5000 TO	24.5000	894	893.02	.00
24.5000 TO	34.5000	707	701.12	.05
34.5000 TO	44.5000	300	298.61	.01
44.5000 TO	54.5000	96	104.06	.62
54.5000 TO	64.5000	32	34.24	.15
64.5000 TO	74.5000	12	11.25	.05
74.5000 TO	84.5000	5	3.78	1.71
84.5000 TO	94.5000	2	1.31	--
94.5000 TO	104.5000	2	.47	--
104.5000 TO		0	.17	--

CHI-SQUARE = 2.61, D.F. = 5, P = .7599

Figure 5.1

BILLICK BLOOD LEAD DATA - BLACKS, AGE 13 - 48 MONTHS, 1972

FIT TO A MIXTURE OF TWO GROUPED LOGNORMAL DISTRIBUTIONS

ESTIMATED VALUE OF MU1 = 3.242488
ESTIMATED VALUE OF MU2 = 3.379345
ESTIMATED VALUE OF SIG1 = .323412
ESTIMATED VALUE OF SIG2 = .492301
ESTIMATED VALUE OF P = .671531

GEOMETRIC MEANS AND STANDARD DEVIATIONS

POPULATION	MEAN	STD. DEV.	SIZE
1	25.59732	1.38183	7315.6
2	29.35154	1.63608	1078.4

ESTIMATE OF THE ASYMPTOTIC COVARIANCE MATRIX

	MU1	MU2	SIG1	SIG2
MU1	.79341-004	-.42596-003	-.12893-004	.23330-003
MU2	-.42596-003	.14216-001	.15778-002	.97512-003
SIG1	-.12893-004	.15778-002	.22145-003	.26874-003
SIG2	.23330-003	.97512-003	.26874-003	.17445-002
P	.11595-004	.10316-001	.14022-002	.24433-002

LOG-LIKELIHOOD = -11654.2288

CHI-SQUARE GOODNESS OF FIT TEST

INTERVAL		OBSERVED	EXPECTED	CHI-SQUARE
TO	14.5000	370	370.40	.00
14.5000 TO	24.5000	3279	3278.08	.00
24.5000 TO	34.5000	3042	3042.63	.00
34.5000 TO	44.5000	1169	1169.01	.00
44.5000 TO	54.5000	350	350.14	.00
54.5000 TO	64.5000	111	108.93	.04
64.5000 TO	74.5000	39	39.77	.02
74.5000 TO	84.5000	16	17.12	.07
84.5000 TO	94.5000	7	8.26	.19
94.5000 TO		11	9.65	.19

CHI-SQUARE = .51, D.F. = 4, P = .9727

Figure 5.2

1972 data does fit well to a mixture of two grouped lognormal distributions. This could result in the mixed distribution seen in figure 5.2. Note that the overall blood lead mean in 1974 was lower than it was in 1972.

6. REFERENCES

Alldredge, J. R., D. D. Bolstad, 1977. GDIST: A Computer Code for Analysis of Statistical Distributions of Physical Data. United States Department of Commerce, NTIS TN33.071, 57.

Ames, B. N., J. McCann, E. Yamasaki, 1975. Method for Detecting Carcinogens and Mutagens with the Salmonella/Mammalian-Microsome Mutagenicity Test. Mutation Research, 31, 347-364.

Ali, M. M., C. Giaccotto, 1982. The Identical Distribution Hypothesis for Stock Market Prices-Location-and Scale-Shift Alternatives. Journal of the American Statistical Association, 77, 19-28.

Billick, I. H., A. S. Curran, D. R. Shier, 1979. Analysis of Pediatric Blood Lead Levels in New York for 1970-1976 Archives of Environmental Health, 31, 180-190.

Birnbaum, Z. W., S. C. Saunders, 1958. A Statistical Model for Life-Length Materials. Journal of the American Statistical Association, 53, 151-159.

BMDP Statistical Software, 1981. University of California Press, Berkeley, California.

Bouver, H., R. E. Bargmann, 1975. Computational Algorithms and Modules for the Evaluation of Statistical Distribution Functions. Univ. of Georgia, Dept. of Statistics, Athens, Georgia.

Day, N. E., 1969. Estimating the Components of a Mixture of Normal Distributions. Biometrika, 56, 3, 463-484.

Dempster, A. D., N. M. Laird, D. B. Rubin, 1977. Maximum Likelihood from Incomplete Data Via the EM Algorithm. Journal of the Royal Statistical Society B, 38, 1-22.

Epstein, B., 1960. Test for the Validity of the Assumption that the Underlying Distribution of Life is Exponential: Part II. Technometrics, 2, 167-183.

Fisher, R. A., 1958. Statistical Methods for Research Workers, Thirteenth Revised Edition, Hafner Publishing Company, Incorporated, New York, New York.

Hartley, H. O., 1961. The Modified Gauss-Newton Method for Fitting of Non-Linear Regression Functions by Least Squares. Technometrics, 3, 269-280.

Haseman, J. K., and E. R. Soares, 1976. The Distribution of Fetal Deaths in Control Mice and Its Implications on Statistical Tests for Dominant Lethal Effects. Mutation Research, 41, 277-288.

Hasselblad, V., 1966. Estimation of Parameters for a Mixture of Normal Distributions. Technometrics, 8, 431-441.

Hasselblad, V., 1969. Estimation of Finite Mixtures of Distributions from the Exponential Family. Journal of the American Statistical Association, 64, 1459-1571.

Hasselblad, V., A. G. Stead, W. Galke, 1980. Applications of Regression Analysis to Grouped Blood Lead Data. Journal of the American Statistical Association, 75, 771-778.

Hasselblad, V., A. G. Stead, H. S. Anderson, 1981. DISFIT: A Distribution Fitting System 1. Discrete Distributions, U. S. Environmental Protection Agency, EPA 600/2-81-010.

Holland, D. M., T. Fitz-Simons, 1982. Fitting Statistical Distributions to Air Quality Data by the Method of Maximum Likelihood. Atmospheric Environment, 15, 1071-1076.

Johnson, N. I., S. Kotz, 1969. Discrete Distributions, John Wiley and Sons, Incorporated, Somerset, New Jersey.

Johnson, N. I., S. Kotz, 1970. Continuous Univariate Distributions-1, Houghton Mifflin Co., Boston, Mass.

Johnson, N. I., S. Kotz, 1970. Continuous Univariate Distributions-2, Houghton Mifflin Co., Boston, Mass.

Kendall, M. G., A. Stuart, 1961. The Advanced Theory of Statistics, Volume 2. Hafner Publishing Company, New York, New York, 39-44.

Kenyon, K. W., V. B. Scheffer, D. G. Chapman, 1954. A Population Study of the Alaska Fur-Seal Herd, U. S. Dept. of the Interior, Washington, D. C.

Morgan, C., W. Nelson, P. Caporal, 1979. General Maximum Likelihood Fitting with STATPAC, General Electric Technical Report.

Orchard, T., M. A. Woodbury, 1972. A Missing Information Principle: Theory and Applications. Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability, 1, 697-715.

SAS User's Guide, 1979. SAS Institute, Cary, North Carolina.

SPSS, Statistical Package for the Social Sciences, 1975. McGraw Hill Book Company, New York, N. Y.

Tepper, L. B., L. S. Levin, 1975. A Survey of Air and Population Lead Levels in Selected American Communities. Environmental Quality and Safety, Supplement II: Lead, 152-195.